UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 3 Tahun 2024 Halaman 279-286.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Classification of Dropout Rates in West Sumatra Using the Random Forest Algorithm with Synthetic Minority Oversampling Technique

Anita Fadila, Syafriandi*, Yenni Kurniawati, and Admi Salma
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: syafriandi_math@fmipa.unp.ac.id

Submitted : 26 Juni 2024
Revised : 12 Agustus 2024
Accepted : 13 Agustus 2024

ABSTRACT

This study aims to classify school dropout rates in West Sumatra Province using the Random Forest algorithm with the Synthetic Minority Oversampling Technique (SMOTE). Based on 2021 data from the Ministry of Education, Culture, Research, and Technology (Kemdikbudristek), the dropout rate in West Sumatra is above the national average. Despite efforts to reduce dropout rates, results remain suboptimal. Therefore, this study seeks to identify the causes of student dropouts and compare the performance of the Random Forest algorithm with and without SMOTE. The study uses the 2021 dropout data from West Sumatra, which has a significant class imbalance. SMOTE is applied to balance the data. The dataset is split into training and testing sets in an 80%:20% ratio, and parameter tuning is performed to optimize mtry and the number of trees (ntree). The model is evaluated using a confusion matrix to compare performance. The results show that Random Forest with SMOTE outperforms the version without SMOTE, with improvements in precision, recall, and F1-score. The presence of the biological mother (X_4) is identified as the most significant factor influencing student dropouts, based on the Mean Decrease Gini value. The study concludes that using SMOTE in the Random Forest algorithm helps reduce classification bias and enhances the model's ability to detect students at risk of dropping out.

Keywords: *Dropout Rates, Random Forest*, *Synthetic Minority Oversampling Technique*

I. PENDAHULUAN

Putus sekolah (*drop out*) secara umum mengacu pada individu atau anak yang meninggalkan pendidikan sebelum menyelesaikan pendidikan sesuai dengan durasi yang ditetapkan oleh sistem pendidikan nasional. Menurut Hikmah (2016), putus sekolah adalah situasi dimana seorang murid tidak mampu menyelesaikan program belajarnya sebelum waktu yang telah ditentukan atau tidak berhasil menamatkan seluruh rangkaian program belajarnya. Menurut Djumhur & Surya (2008), terdapat tiga jenis putus sekolah, yaitu berhenti di tengah jenjang pendidikan, berhenti di akhir jenjang pendidikan, dan berhenti antara dua tingkatan pendidikan. Berdasarkan data Kementrian Pendidikan, Kebudayaan, Riset dan Teknologi (Kemdikbudristek) tahun 2021, rata-rata angka putus sekolah di Sumatera Barat sebesar 0,37% yang berada di atas rata-rata nasional yakni sebesar 0,28%. Di tahun 2021, angka putus sekolah di Provinsi Sumatera Barat pada jenjang pendidikan Sekolah Dasar (SD) adalah sebesar 0,11%, jenjang Sekolah Menengah Pertama (SMP) sebesar 0,60%, dan pada jenjang Sekolah Menengah Atas sebesar 0,76%.

Berbagai upaya telah dilakukan oleh pemerintah pusat untuk mengatasi masalah putus sekolah, terutama yang disebabkan oleh masalah ekonomi. Dengan pemberian program-program seperti Bantuan Operasional Sekolah (BOS), Bantuan Siswa Miskin (BSM), Program Keluarga Harapan (PKH), Program Indonesia Pintar (PIP), dan berbagai beasiswa dari pemerintah pusat maupun daerah juga telah diluncurkan. Namun, solusi-solusi tersebut belum mampu sepenuhnya mengatasi masalah putus sekolah. Hal ini menunjukkan bahwa di Provinsi Sumatera Barat, tidak hanya permasalahan ekonomi yang menjadi penyebab putus sekolah. Diduga ada faktor lain yang turut mempengaruhi angka putus sekolah, seperti pendidikan orang tua, jumlah bersaudara, lokasi tempat tinggal, dan lain-lain. Oleh karena itu, diperlukan penelitian lebih lanjut untuk mengidentifikasi penyebab siswa putus sekolah, sehingga pemerintah dan masyarakat dapat berperan lebih efektif dalam menekan angka putus sekolah. Penerapan statistika seperti metode klasifikasi dalam *machine learning*, dapat dimanfaatkan untuk memprediksi atribut yang menjadi penyebab siswa putus sekolah. Klasifikasi dapat membantu mengidentifikasi dan memprediksi kelompok-kelompok yang rentan mengalami putus sekolah. Sebagai contoh, anak-anak dari keluarga dengan tingkat pendidikan rendah mungkin lebih rentan untuk putus sekolah dibandingkan anak-anak dari keluarga dengan tingkat pendidikan yang lebih tinggi. Dengan demikian, program-program atau solusi yang ditawarkan dapat diarahkan secara lebih efektif kepada kelompok-kelompok yang membutuhkan.

Random forest adalah salah satu metode klasifikasi yang populer karena kemampuannya dalam membuat banyak pohon keputusan dan menggabungkan hasil prediksinya untuk menghasilkan prediksi akhir yang lebih akurat (Khalilia

DOI: https://doi.org/10.24036/ujsds/vol2-iss3/183

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

dkk., 2011). Kelebihan lain dari metode ini adalah *Random Forest* dapat bekerja dengan baik pada data yang memiliki banyak atribut tanpa memerlukan seleksi secara manual, bekerja dengan baik baik pada dataset yang seimbang maupun tidak seimbang, dan memberikan estimasi atribut penting dalam proses (Juarez dkk., 2018). Namun, menurut Erlin dkk. (2022), sebagian besar teknik pemodelan, termasuk *random forest*, bekerja paling baik ketika distribusi kelas dalam dataset seimbang. Pada kenyataannya, banyak dataset yang tidak seimbang di mana kelas mayoritas jauh lebih banyak dibandingkan kelas minoritas. Hal ini menyebabkan model cenderung bias dan memiliki akurasi rendah dalam memprediksi kelas minoritas (Jian dkk., 2016).

Untuk mengatasi masalah data tidak seimbang ini, pendekatan yang dapat digunakan adalah metode *Synthetic Minority Oversampling Technique* (SMOTE) yaitu menambah data pada kelas minoritas sehingga sebaran data menjadi lebih seimbang tanpa kehilangan informasi penting pada data (Ramentol dkk., 2012). Penelitian oleh Michael dkk. (2023), menunjukkan bahwa kombinasi metode *random forest* dengan SMOTE untuk klasifikasi kanker paru-paru meningkatkan kinerja prediksi secara signifikan, dengan akurasi 94.1%, sensitivitas 94.5%, dan spesifisitas 93.7%. Sebagai perbandingan, *random forest* tanpa SMOTE hanya menghasilkan akurasi 89.1%, sensitivitas 55%, dan spesifisitas 94.5%.

Oleh karena itu, pada penelitian ini dilakukan klasifikasi angka putus sekolah menggunakan algortima *random forest* dengan SMOTE di Provinsi Sumatera Barat Tahun 2021. Penelitian ini dapat membantu lembaga pemerintahan mengidentifikasi atribut yang paling berkontribusi terhadap angka putus sekolah. Dengan begitu pemerintah dapat merancang kebijakan dan program intervensi yang lebih efektif untuk mencegah putus sekolah.

II. METODE PENELITIAN

Jenis penelitian ini adalah penelitian terapan. Penerapan algoritma *random forest* dengan SMOTE dalam mengklasifikasikan angka putus sekolah di Provinsi Sumatera Barat tahun 2021. Data yang digunakan adalah data putus sekolah diperoleh dari hasil Survey Sosial Ekonomi Nasional (Susenas) Modul Sosial, Budaya, dan Pendidikan (MSBP) Badan Pusat Statistik (BPS) di Provinsi Sumatera Barat tahun 2021 yang terdiri atas 2235 amatan. Variabel penelitian terbagi atas variabel respon (label) dan variabel bebas (atribut) yang disajikan dalam Tabel 1.

Tabel 1. Variabel Penelitian

No	Variabel	Skala	Kategori
1	Status putus sekolah (Y)	Nominal	0 = Tidak putus sekolah dan 1 = Putus sekolah
2	Jenis Kelamin (X_1)	Nominal	0 = Perempuan dan 1 = Laki-Laki
3	Lokasi Sekolah (X_2)	Nominal	0 = Pedesaan dan 1 = Perkotaan
4	Pengawasan Belajar (X_3)	Nominal	0 = Tidak, 1 = Ya
5	Keberadaan Ibu Kandung (X_4)	Nominal	0 = Meninggal dan 1 = Hidup
6	Keberadaan Ayah Kandung (X_5)	Nominal	0 = Meninggal dan 1 = Hidup
7	Kelengkapan Fasilitas Belajar (X_6)	Nominal	0 = Tidak dan 1 = Ya
8	Jumlah Anggota Rumah Tangga (X)	Nominal	0 = ART $\leq$ 4 dan 1 = ART $>$ 4
9	Penerima Beasiswa (X_8)	Nominal	0 = Tidak, 1 = Ya
10	Kategori Pembayaran SPP (X_9)	Nominal	0 = SPP Rendah < Rp223.516,- dan 1 = SPP Tinggi $\geq$ Rp223.516,-
11	Kategori Pembayaran Komite (X_{10})	Nominal	0 = Komite Rendah < Rp64.439,- dan 1 = Komite Tinggi $\geq$ Rp64.439,-

A. Tahapan Analisis Data

Tahapan analisis data sebagai berikut.

1. Melakukan pengambilan data yang diperlukan, yaitu data Susenas MSBP diperoleh dari Badan Pusat Statistik Provinsi Sumatera Barat tahun 2021.
2. *Preprocessing* data dengan melakukan pelabelan data pada peubah sesuai dengan label yang ditetapkan dan melakukan visualisasi data untuk melihat proporsi data.
3. Mengatasi ketidakseimbangan data dengan metode SMOTE.

   Menurut Chawla dkk., (2002), *Synthetic Minority Oversampling Technique* (SMOTE) dianggap sebagai solusi untuk mengatasi data tidak seimbang, dimana teknik SMOTE menambah amatan kelas minoritas agar setara dengan kelas mayoritas dengan cara membangkitkan data sintetis (data buatan) berdasarkan prinsip tetangga terdekat menggunakan *K-Nearest Neighbor* (KNN). Menurut Wijaya dkk., (2018), SMOTE dapat meningkatkan nilai *sensitivity* yang lebih dari 50%, meskipun nilai *accuracy* dan *specificity* menurun dibandingkan dengan model sebelum SMOTE. Cara kerja SMOTE sebagai berikut.

   a. Menetapkan kelas minoritas.
   b. Menentukan jarak k-tetangga terdekat (KNN) dari X yang diperoleh dengan menghitung jarak *Value Difference Metric* (VDM) menggunakan persamaan berikut.

DOI: https://doi.org/10.24036/ujsds/vol2-iss3/183

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

$$d(x_i, y_i) = \sum_{j=1}^{k} \left| \frac{C_{xi}}{C_x} - \frac{C_{yi}}{C_y} \right|$$

Dengan k adalah jumlah kelas dalam data, C_{xi}, C_{yi} adalah jumlah jarak dengan nilai atribut x dan y yang termasuk ke dalam kelas ke-i. C_x, C_y adalah jumlah total jarak dengan nilai atribut x dan y.

c. Membuat sampel sintetis baru dengan menggunakan persamaan berikut.

$$S = x + u.(x^R - x) \quad (1)$$

dimana S adalah sampel sintetis baru, x adalah sampel kelas minoritas, u adalah bilangan acak bernilai [0,1], dan x^R adalah tetangga terdekat dari sampel kelas minoritas

4. Melakukan klasifikasi menggunakan algoritma *random forest.*

Menurut Breiman dkk., (2017), *random forest* adalah pengembangan dari metode *Classification and Regression Tree* (CART). Metode *random forest* berbeda dengan metode CART yang berusaha mencari output maksimal dalam sekali percobaan, sedangkan *random forest* merupakan kumpulan banyak *decision tree* untuk membangun satu pohon yang mungkin menghasilkan prediktif yang lebih akurat. Menurut Mishina dkk., (2015), *random forest* adalah algoritma pembelajaran mesin yang menggabungkan *bootstrap aggregating* (*bagging*) dan pemilihan fitur untuk memperkenalkan keacakan dan mudah diparalelkan. Konsep dasar *random forest* berasal dari metode *bagging* dengan memilih fitur secara acak untuk mengurangi korelasi antar pohon yang dibentuk. Langkah klasifikasi menggunakan algoritma *random forest* sebagai berikut.

a. Membagi dataset menjadi data *traning* dan data *testing*.

Menurut Gorunescu (2011), data *training* adalah kumpulan data yang berisi nilai-nilai dari kelas dan predictor yang berfungsi untuk membuat model. Sedangkan data *testing* merupakan sekumpulan data yang digunakan untuk proses pengujian dalam klasifikasi terhadap model yang dihasilkan sebelumnya serta memungkinkan untuk melakukan proses evaluasi akurasi klasifikasi terhadap hasil. Menurut Baiq dkk., (2023), perbandingan proporsi data *training* dan *testing* bisa beragam tergantung pada ukuran dataset dan kebutuhan analisis. Dalam penelitian ini perbandingan data *training* dan *testing* yang digunakan adalah 80%:20%.

b. Membangun pohon ke-i dari data training atau dalam random forest disebut sampel *bootstrap*. *Bootstrap* merupakan suatu metode berbasis resampling data dengan syarat pengembalian. Pada tahap ini perlu ditentukan nilai *mtry* yang digunakan, yaitu $\sqrt{p}$ dengan *mtry* < p (variabel bebas). Selanjutnya menghitung nilai impurity simpul kanan dan kiri dan menghitung perhitungan *gini impurity/gini index* menggunakan persamaan berikut.

Node kiri (L) : $imp(t_L) = \sum_{l=1}^{2} p_{t_L}^{(l)}\left(1 - p_{t_L}^{(l)}\right)$ (2)

Node kanan (R) : $imp(t_R) = \sum_{l=1}^{2} p_{t_R}^{(l)}\left(1 - p_{t_R}^{(l)}\right)$ (3)

$$imp(t) = \sum_{k=1}^{2} p_t^{(k)}\left(1 - p_t^{(k)}\right) \quad (4)$$

dengan :

$p_t^{(l)} = \frac{n_t^{(l)}}{n_t}$,$p_t^{(k)} = \frac{n_t^{(k)}}{n_t}$ dimana $p_t^{(l)}$,$p_t^{(k)}$ adalah perbandingan objek kelas ke-i dan ke-k pada *node t*, $n_t^{(l)}$ dan$n_t^{(k)}$ adalah jumlah observasi kelas ke-i dan ke-k pada *node t*, dan $n_t$adalah jumlah seluruh observasi pada node t

Lalu menghitung *Goodness of Split* untuk pemilah variabel *s* pada *node t* menggunakan persamaan berikut.

$$\Delta imp(s,t) = imp(t) - p_{t_L} imp(t_L) - p_{t_R} imp(t_R) \quad (5)$$

dengan,

$p_{t_L} = \frac{n_{t_L}}{n_t}$ dan $p_{t_R} = \frac{n_{t_R}}{n_t}$ dimana p_{t_L}, p_{t_R} adalah perbandingan objek pada *node t* yang memilah pada *node* t_L atau t_R dan n_{t_L}. n_{t_R} adalah jumlah observasi pada *node t* yang memilah pada *node* t_L atau t_R

c. Setelah pohon dan *forest* terbentuk, selanjutnya menghitng tingkat misklasifikasi menggunakan sampel *Out of Bag* (OOB) pada masing-masing kombinasi *mtry* dan *ntree*. Berikut persamaan untuk menghitung nilai laju galat pada satu pohon.

$$Laju\ Galat\ OOB_i = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_{i \neq \hat{Y}_i}}$$

DOI: https://doi.org/10.24036/ujsds/vol2-iss3/183

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

dimana, $\sum_{i=1}^{n} 1_{Y_{i \neq \hat{Y}_i}}$: jumlah data hasil misklasifikasi, Y_i adalah hasil amatan aktual ke-i, $\hat{Y}_i$ adalah hasil amatan prediksi ke-i dan n adalah jumlah OOB ke-i yang menjadi sampel OOB

Setelah memperoleh nilai tingkat misklasifikasi dari sampel OOB ke-i, maka dilanjutkan dengan menghitung rataan tingkat misklasifikasi OOB dengan persamaan berikut.

$$Laju\ Galat\ OOB = \frac{\sum OOB_i\ laju\ galat}{k} x100\% \quad (6)$$

d. Menentukan penerapan yang paling optimal berdasarkan tunning parameter *mtry* dan *ntree* yang dikombinasikan berdasarkan nilai total *error rate* OOB terkecil.

e. Mengidentifikasi atribut penting.
Tingkat kepentingan peubah penjelas (*important variable*) yang dihasilkan oleh *random forest* dapat dilihat dengan melakukan perhitungan pengukuran *variable importance measure* (VIM) dengan *Mean Decrease Gini* (MDG). Misalkan terdapat p peubah penjelas dengan m pohon, MDG mengukur tingkat kepentingan peubah penjelas X dengan cara berikut (Breiman, 2001).

$$MDG = \frac{1}{K} \sum_t [\Delta i(s,t) I(s,t)] \quad (7)$$

dimana,
K : banyaknya pohon yang terbentuk
$\Delta i(s,t)$: nilai impuritas tereduksi untuk peubah penjelas X_s pada simpul t
$I(s,t)$: fungsi indikator yang bernilai 1 ketika X_s memilah simpul t dan 0 lainnya

5. Melakukan evaluasi model menggunakan *confussion matrix*.
Menurut Han dan Kamber (2006), *confussion matrix* adalah alat yang digunakan dalam evaluasi kinerja model klasifikasi di *machine learning*. *Confussion matrix* disajikan dalam bentuk tabel yang menggambarkan kinerja model klasifikasi dengan membandingkan prediksi model terhadap nilai sebenarnya dari data uji. Komponen dalam *confusion matrix* memiliki dua kelas dengan empat kemungkinan hasil prediksi klasifikasi yang berbeda yang disajikan pada Tabel 2.

Tabel 2. *Confusion Matrix*

Classification		*Predicted Value*	
		Class = Positive	*Class = Negative*
Actual Value	*Class = Positive*	*True Positif (TP)*	*False Positif (FP)*
	Class = Negative	*False Negatif (FN)*	*True Negatif (TN)*

Dari Tabel 2, dapat dihitung nilai berikut.

a. *Accuracy* digunakan untuk mengukur sejauh mana model klasifikasi berhasil dalam memprediksi secara akurat semua label. *Accuracy* dapat dihitung dengan persamaan berikut.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (8)$$

b. *Precision* merupakan perbandingan prediksi yang benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. *Precision* dapat dihitung dengan persamaan berikut.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (9)$$

c. *Recall* atau *sensitivity* merupakan perbandingan prediksi yang benar positif dibandingkan dengan keseluruhan data yang benar positif. *Recall* dapat dihitung dengan persamaan berikut.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (10)$$

d. *Specificity* merupakan perbandingan kebenaran memprediksi yang negatif dibandingkan dengan keseluruhan data negatif. *Specificity* dapat dihitung dengan persamaan berikut.

$$Specificity = \frac{TN}{TN + FP} \quad (11)$$

e. *F1-Score* merupakan gambaran keakuratan model dalam mengklasifikasikan dengan benar. F1-*score* memberikan perbandingan yang seimbang antara antara *precision* dan *recall* dan berguna ketika memiliki data dengan distribusi kelas yang tidak seimbang. *F1-Score* dapat dihitung dengan persamaan berikut.

$$\text{F1-}Score = 2 * \frac{precision * recall}{precision + recall} \quad (12)$$

DOI: https://doi.org/10.24036/ujsds/vol2-iss3/183

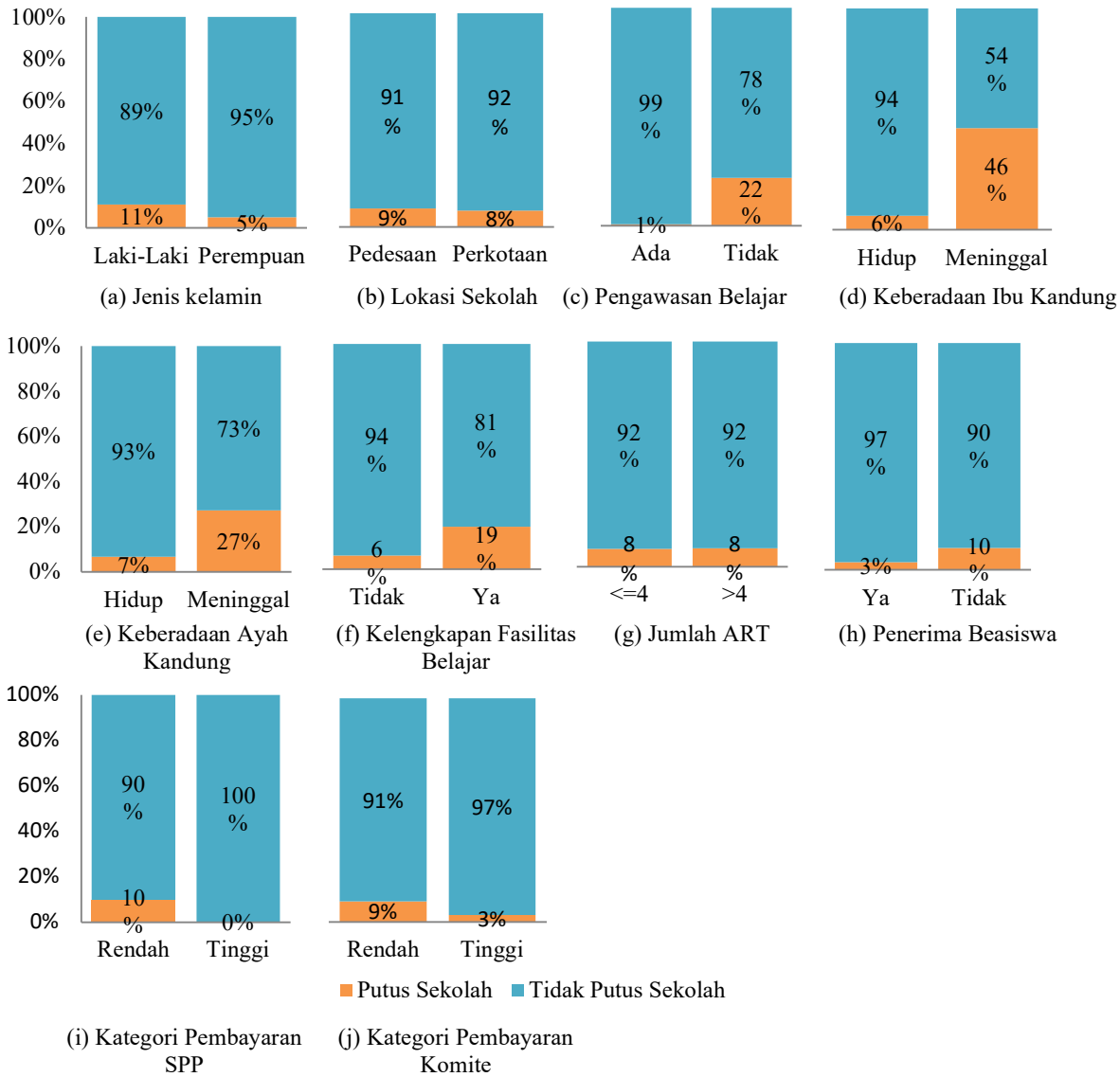
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

6. Melihat perbandingan nilai *accuracy*, *precision*, *recall, specificity* dan F1-*score* dari klasifikasi algoritma *random forest* sebelum SMOTE dan dengan SMOTE.
7. Menarik kesimpulan.

III. HASIL DAN PEMBAHASAN

A. VISUALISASI DATA

Analisis data dilakukan menggunakan *software R Studio*. Sebelum melakukan analisis, langkah pertama yang dilakukan adalah melakukan visualisasi data. Visualisasi data digunakan untuk melihat proporsi data sebelum proses penyeimbangan data yang dilakukan dengan metode SMOTE. Gambar 1 menunjukkan visualisasi data berdasarkan karakteristik angka putus sekolah berdasarkan atribut.



Gambar 1. Persentase Putus Sekolah dan Tidak Putus Sekolah Berdasarkan Atribut

Visualisasi Gambar 1 menunjukkan bahwa atribut yang dapat membedakan antara siswa yang putus sekolah dan tidak putus sekolah dilihat dari persentase putus sekolah yang tinggi adalah atribut keberadaan ibu kandung, keberadaan ayah kandung, dan pengawasan belajar.

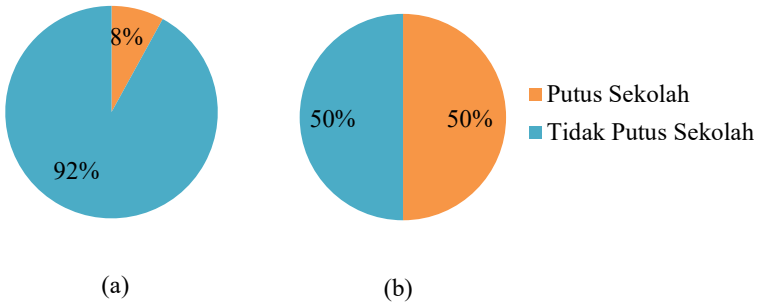
B. *Synthetic Minority Oversampling Technique*

Visualisasi data pada Gambar 2 menunjukkan bahwa proporsi data antara putus sekolah dan tidak putus sekolah tidak seimbang, dimana amatan cenderung berada di kelas tidak putus sekolah. Keadaan ini menyebabkan kelas putus

DOI: https://doi.org/10.24036/ujsds/vol2-iss3/183

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

sekolah menjadi kelas minoritas dan kelas tidak putus sekolah menjadi kelas mayoritas. Sehingga sebelum melakukan analisis data harus diseimbangkan terlebih dahulu menggunakan metode SMOTE dengan Persamaaan (1). Gambar 2 menunjukkan visualisasi label sebelum dan sesudah dilakukan SMOTE.



(a) (b)

Gambar 2. (a) Label Sebelum Diseimbangkan dan (b) Label Setelah Diseimbangkan

Gambar 2 (a) menunjukkan bahwa proporsi kelas putus sekolah sangat kecil dibanding yang tidak putus sekolah. Kelas putus sekolah menjadi kelas minoritas dengan persentase sebesar 8% atau amatan sebanyak 183, sedangkan kelas tidak putus sekolah menjadi kelas mayoritas dengan persentase sebesar 92% atau amatan sebanyak 2052. Setelah dilakukan SMOTE pada data, diperoleh data yang seimbang yang ditunjukkan pada Gambar 2 (b), dimana proses penyeimbangan data menghasilkan proporsi yang sama antara kelas putus sekolah dan tidak putus sekolah. Persentase kelas putus sekolah menjadi sebesar 50% dengan amatan sebanyak 2052 dan persentase kelas tidak putus sekolah adalah 50% dengan amatan sebanyak 2052.

C. *Random Forest*

Random forest merupakan kumpulan dari banyak pohon, dimana proses pembentukan pohon dilakukan menggunakan Persamaan (2), (3), (4) dan (5). Parameter *mtry* yang dipakai yaitu $\sqrt{p} \approx 3$ dengan kombinasi *ntree* yaitu 100, 250, 500 dan 1000. Tabel 3 menunjukkan nilai hasil dari *tuning* parameter yang tepat akan menghasilkan forest yang optimal. Semakin kecil nilai laju galat OOB, maka semakin sedikit misklasifikasi pada *forest* yang berarti semakin baik *forest* yang terbentuk. Nilai laju galat OOB terhadap kasus angka putus sekolah di Sumatera Barat tahun 2021 dapat dilihat pada Tabel 3.

Tabel 3. *Tuning* Parameter *Random Forest*

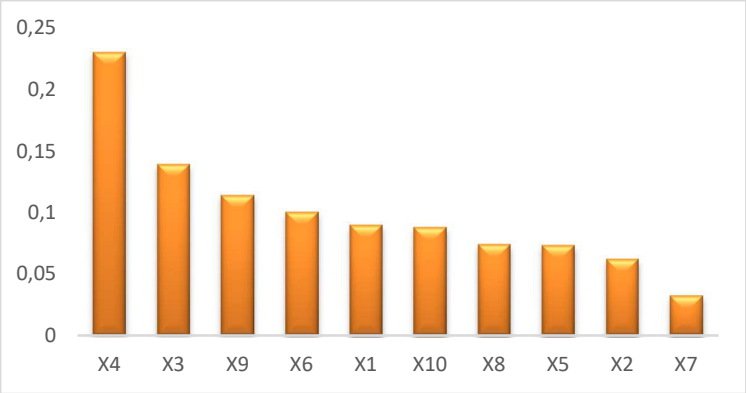
Algoritma	Laju Galat OOB			
	Ntree=100	*Ntree*=250	*Ntree*=500	*Ntree*=1000
Random Forest tanpa SMOTE	6,71%	6,71%	**6,67%**	6,76%
Random Forest dengan SMOTE	12,76%	12,70%	**12,67%**	12,70%

Pada Tabel 3 terlihat bahwa penerapan *random forest* pada klasifikasi angka putus sekolah, baik tanpa SMOTE atau dengan SMOTE terlihat bahawa *forest* yang optimal yaitu *forest* dengan kombinasi *mtry* = 3 dan *ntree* = 500. Nilai OOB *error rate* ini berguna untuk melihat tingkat misklasifikasi *forest* dalam mengklasifikasikan status putus sekolah dan tidak putus sekolah di Provinsi Sumatera Barat tahun 2021. Selain melihat tingkat misklasifikasi *forest*, hasil kombinasi ini nantinya akan digunakan dalam proses lanjutan evaluasi model menggunakan *confussion matrix.*

Berdasarkan hasil penerapan *random forest* yang optimal, selanjutnya melihat kepentingan atribut pada data atau VIM. VIM dilihat dari nilai rataan penurunan indeks gini (MDG) yang diperoleh dari *forest* yang terbentuk. VIM menunjukkan atribut mana saja yang berpengaruh dalam angka putus sekolah. Nilai MDG dari masing-masing kombinasi *mtry* dan *ntree* yang dicobakan disajikan pada Gambar 3.

DOI: https://doi.org/10.24036/ujsds/vol2-iss3/183

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



Gambar 3. VIM Angka Putus Sekolah

Gambar 3 menunjukkan bahwa tingkat kepetingan atribut diurutkan dari terbesar hingga terkecil. Urutan atribut yang paling penting adalah keberadaan ibu kandung ($\boldsymbol{X_4}$), dilanjutkan dengan atribut lainnya yaitu pengawasan belajar ($\boldsymbol{X_3}$), kategori pembayaran SPP ($\boldsymbol{X_9}$), kelengkapan fasilitas belajar ($\boldsymbol{X_6}$), jenis kelamin ($\boldsymbol{X_1}$), kategori pembayaran komite ($\boldsymbol{X_{10}}$), penerima beasiswa ($\boldsymbol{X_8}$), keberadaan ayah kandung ($\boldsymbol{X_5}$), lokasi sekolah ($\boldsymbol{X_2}$), dan jumlah ART ($\boldsymbol{X_7}$). Hal ini menunjukkan bahwa dalam kasus putus sekolah keberadaan ibu kandung hidup atau meninggal menjadi pertimbangan utama yang memutuskan apakah seseorang akan putus sekolah atau tidak dibanding atribut-atribut lainnya

D. *Confussion Matrix*

Metode yang digunakan dalam evaluasi model adalah *confussion matrix* dengan menggunakan data *testing*. Data testing tanpa SMOTE terdapat 447 data dari 2235 data keseluruhan, sedangkan data testing dengan SMOTE terdapat 821 data dari 4104 data keseluruhan. *Confusion matrix* dari algoritma *random forest* dengan SMOTE dan sebelum SMOTE disajikan pada Tabel 4.

Tabel 4. Nilai *Confussion Matrix Random Forest*

Algoritma	***Confusion Matrix***			
	True Positive	***False Positive***	***True Negative***	***False Negative***
Random Forest tanpa SMOTE	13	5	404	25
Random Forest dengan SMOTE	356	80	348	37

Berikut merupakan perhitungan nilai *accuracy, precision, recall*, *specificity* dan F1-*score* dari klasifikasi angka putus sekolah menggunakan algoritma *random forest* dengan SMOTE dan sebelum SMOTE disajikan pada Tabel 5.

Tabel 5. Nilai *Accuracy, Precision, Recall*, dan *Specificity*

Kinerja	***Random Forest* tanpa SMOTE**	***Random Forest* dengan SMOTE**
Accuracy	93%*	86%
Precision	72%	82%*
Recall (*Sensitivity*)	34%	91%*
Specificity	99%*	81%
F1-Score	46%	86%*

Tabel 5 menunjukkan bahwa model tanpa SMOTE menghasilkan nilai *accuracy* dan *specificity* yang lebih tinggi dibanding dengan SMOTE. Sedangkan untuk *precision, recall*, dan F1-*score*, model dengan SMOTE memberikan nilai yang lebih tinggi dibanding tanpa SMOTE. Artinya, model dengan SMOTE lebih baik dalam melakukan klasifikasi angka putus sekolah karena mampu mengatasi bias dalam klasifikasi yang cenderung memprediksi kelas mayor (tidak putus sekolah) dibanding kelas minor (putus sekolah).

IV. KESIMPULAN

Algoritma *random forest* menghasilkan model yang optimal untuk mengidentifikasi angka putus sekolah di Sumatera Barat tahun 2021 dengan menggunakan kombinasi *tuning* parameter *mtry* = 3 dan *ntree* = 500. Berdasarkan hasil forest optimal diperoleh atribut keberadaan ibu kandung menjadi atribut utama yang menentukan seseorang putus sekolah atau tidak. Jika dilihat dari akurasi model tanpa SMOTE memberikan akurasi yang lebih besar dibanding model dengan SMOTE, hal ini terjadi karena hasil klasifikasi cenderung memprediksi kelas mayor dibanding minor. Namun model dengan SMOTE mampu untuk mengklasifikasikan kategori kelas minor yaitu putus sekolah dengan baik, dilihat

DOI: https://doi.org/10.24036/ujsds/vol2-iss3/183

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

dari peningkatan nilai persentase *presisi*, *recall* dan F1-*score*. Sehingga model dengan SMOTE dapat menangani data dengan lebih baik dan memberikan prediksi yang lebih andal.

UCAPAN TERIMA KASIH

Terima kasih kepada instansi penyedia data dalam penelitian ini yaitu Badan Pusat Statistik (BPS) Sumatera Barat.

DAFTAR PUSTAKA

Badan Pusat Statistik Provinsi Sumatera Barat. (2021). *Profil Pendidikan Provinsi Sumatera Barat 2021*. Sumatera Barat : Badan Pusat Statistik.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification And Regression Trees*. Routledge. Chapman and Hall : New York. https://doi.org/10.1201/9781315139470

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Erlin, E., Desnelita, Y., Nasution, N., Suryati, L., & Zoromi, F. (2022). Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *21*(3), 677–690. https://doi.org/10.30812/matrik.v21i3.1726

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*. New York : Springer

Han J, & Kamber M. (2006). *Data Mining: Concepts and Techniques*. San Francisco : Morgan Kaufmann.

Hikmah, L. (2016). Kemiskinan dan Putus Sekolah. *Equilibrium Pendidikan Sosiologi* , *4*(2), 2339–2401.

Jian, C., Gao, J., & Ao, Y. (2016). A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing*, *193*, 115–122. https://doi.org/10.1016/j.neucom.2016.02.006

Juarez, O. L. E., Martinez, M. O., Nesterov, S. V., Kajander, S., & Knuuti, J. (2018). The machine learning horizon in cardiac hybrid imaging. *European Journal of Hybrid Imaging*, *2*(1), 15. https://doi.org/10.1186/s41824-018-0033-3

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, *11*(1), 51. https://doi.org/10.1186/1472-6947-11-51

Mishina, Y., Murata, R., Yamauchi, Y., Yamashita, T., & Fujiyoshi, H. (2015). Boosted Random Forest. *IEICE Transactions on Information and Systems*, *E98.D*(9), 1630–1636. https://doi.org/10.1587/transinf.2014OPP0004

Ramentol, E., Caballero, Y., Bello, R., & Herrera, F. (2012). SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowledge and Information Systems*, *33*(2), 245–265. https://doi.org/10.1007/s10115-011-0465-6

Wijaya, J., Soleh, A. M., & Rizki, A. (2018). Penanganan Data Tidak Seimbang Pada Pemodelan Rotation Forest Keberhasilan Studi Mahasiswa Program Magister IPB. Xplore, 2(2), 32–40.

DOI: https://doi.org/10.24036/ujsds/vol2-iss3/183